Information Retrieval: Personalization

Fernando Diaz
Microsoft Research NYC

November 7, 2016
Outline

Introduction to Personalization

Topic-Specific PageRank

News Personalization

Deciding When To Personalize
Introduction to Personalization
given a **query** and a **corpus**, find **relevant documents**.

- **query**: user’s expression of the information need
- **corpus**: the repository of retrievable items
- **relevance**: satisfaction of the information need
given a **query** and a **corpus**, find **relevant documents**.

- **query**: user’s expression of the information need *as well as the context*
- **corpus**: the repository of retrievable items
- **relevance**: satisfaction of the information need *for this user*
User Context

- Task context: the broader task in which this query is occurring
  - shopping
  - research
  - ...
- Group context: demographic factors of the user which may influence relevance
  - location
  - age
  - ...
- Individual context: information particular to this user
  - historic view/click information
  - ...
Explicit versus Implicit Personalization

- Explicit personalization
  - User explicitly defines their topical interests.
  - Advantages:
    - allows a user fine-grained control over personalization.
  - Disadvantages:
    - requires defining a taxonomy/interface for feedback.
    - users may provide disingenuous personalization information.
    - representation is limited by the feedback taxonomy.
Explicit versus Implicit Personalization

- Implicit personalization
  - The system infers topical interests from user behavior.
  - Advantages:
    - allows a rich, data-driven representation.
    - no need to define taxonomy/interface.
  - Disadvantages:
    - user has no explicit feedback mechanism.
    - requires good modeling.
Client-Side Personalization

- Advantages
- Disadvantages
Client-Side Personalization

- Advantages
  - preserves privacy.
  - richer user signals.
- Disadvantages
Client-Side Personalization

- Advantages
  - preserves privacy.
  - richer user signals.
- Disadvantages
  - no access to data from other users/groups.
  - network latency for interfacing with the index (potentially also subject to query quotas).
  - client space/processing requirements.
Server-Side Personalization

- Client-side personalization is inherently decentralized.
  - evaluation requires a controlled experiment similar to a user study.
  - testing new systems/algorithms difficult.
- Server-side personalization centralizes data collection, allowing \textit{in situ} evaluation.
  - advantages?
  - disadvantages?
Today

- Mechanisms for personalization
  - Ranking: topic-specific PageRank
  - Filtering: Google news
- Deciding when to personalize
Topic-Specific PageRank
Topic Specific PageRank

- PageRank
  - Models how often a web page is visited by users.
  - Assumes a surfer randomly navigating between pages.
  - Effective model when user tracking data is missing.
- Topic-Specific PageRank
  - Models how often a web page is visited by users interested in a specific topic.
  - Assumes a surfer interested in a specific topic randomly navigating between pages.
Random Surfer Model
[Brin and Page 1998]

- Simulate a very large number of users browsing the entire web.
- Let users browse randomly. This is a naïve assumption but works okay in practice.
- Observe how often pages get visited.
- The authoritativeness of a page is a function of its popularity in the simulation.
Random Surfer Model
[Brin and Page 1998]

- The matrix, G defines a transition matrix over the web graph.
- In order to run the simulation, we take the matrix-vector product,

\[ G \rho = G \times \begin{bmatrix} \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{6} \end{bmatrix} \]

- The result is a distribution over graph nodes representing where users would have gone have a single simulation step.
Biased Surfer Model

[Haveliwala 2002]

- Assume the surfer does not restart from a random state but that have a *biased* probability of landing in each of the nodes.

\[
p^{\text{sports}} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} \text{sports} \\ \text{arts} \\ \text{politics} \\ \text{business} \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} \frac{1}{2} \\ 0 \\ 0 \\ 1/2 \end{bmatrix}
\]

\[
p^{\text{art}} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} \text{sports} \\ \text{arts} \\ \text{politics} \\ \text{business} \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 0 \end{bmatrix}
\]
Biased Surfer Model

[Haveliwala 2002]

- We can run the simulation for \( t \) step for each topic,

\[
p_{t+1}^{\text{sports}} = G^t p^{\text{sports}}
\]

- The result of this simulation is the topic-based PageRank score for every document in the graph.
Query-Based Ranking
[Haveliwala 2002]

- For a given query, $q$, we compute the topical distribution as,

$$P(c_j|q) = \frac{P(c_j)P(q|c_j)}{P(q)}$$
Query-Based Ranking

[Haveliwala 2002]

- For a given query, $q$, we compute the topical distribution as,

$$P(c_j|q) = \frac{P(c_j)P(q|c_j)}{P(q)} \propto P(c_j)P(q|c_j)$$
Query-Based Ranking

[Haveliwala 2002]

- For a given query, $q$, we compute the topical distribution as,

$$P(c_j | q) = \frac{P(c_j) P(q | c_j)}{P(q)}$$

$$\propto P(c_j) P(q | c_j)$$

$$\propto P(c_j) \prod_{i} P(q_i | c_j)$$
For a given query, $q$, we compute the topical distribution as,

$$P(c_j|q) = \frac{P(c_j)P(q|c_j)}{P(q)}$$

$$\propto P(c_j)P(q|c_j)$$

$$\propto P(c_j) \prod_{i} P(q_i|c_j)$$

Where does $P(q_i|c_j)$ come from?
Query-Based Ranking

[Haveliani 2002]

- For a given query, $q$, we rank documents according to,

$$s_{qd} = \sum_j P(c_j|q) \text{rank}_{jd}$$
Experiments
[Haveliwala 2002]

Table 1: Queries used

<table>
<thead>
<tr>
<th>affirmative action</th>
<th>lipari</th>
</tr>
</thead>
<tbody>
<tr>
<td>alcoholism</td>
<td>lyme disease</td>
</tr>
<tr>
<td>amusement parks</td>
<td>mutual funds</td>
</tr>
<tr>
<td>architecture</td>
<td>national parks</td>
</tr>
<tr>
<td>bicycling</td>
<td>parallel architecture</td>
</tr>
<tr>
<td>blues</td>
<td>recycling cans</td>
</tr>
<tr>
<td>cheese</td>
<td>rock climbing</td>
</tr>
<tr>
<td>citrus groves</td>
<td>san francisco</td>
</tr>
<tr>
<td>classical guitar</td>
<td>shakespeare</td>
</tr>
<tr>
<td>computer vision</td>
<td>stamp collecting</td>
</tr>
<tr>
<td>cruises</td>
<td>sushi</td>
</tr>
<tr>
<td>death valley</td>
<td>table tennis</td>
</tr>
<tr>
<td>field hockey</td>
<td>telecommuting</td>
</tr>
<tr>
<td>gardening</td>
<td>vintage cars</td>
</tr>
<tr>
<td>graphic design</td>
<td>volcano</td>
</tr>
<tr>
<td>gulf war</td>
<td>zen buddhism</td>
</tr>
<tr>
<td>hiv</td>
<td>zener</td>
</tr>
<tr>
<td>java</td>
<td></td>
</tr>
</tbody>
</table>
Experiments
[Haveliwala 2002]

Table 1: Queries used

<table>
<thead>
<tr>
<th>affirmative action</th>
<th>lipari</th>
</tr>
</thead>
<tbody>
<tr>
<td>alcoholism</td>
<td>lyme disease</td>
</tr>
<tr>
<td>amusement parks</td>
<td>mutual funds</td>
</tr>
<tr>
<td>architecture</td>
<td>national parks</td>
</tr>
<tr>
<td>bicycling</td>
<td>parallel architecture</td>
</tr>
<tr>
<td>blues</td>
<td>recycling cans</td>
</tr>
<tr>
<td>cheese</td>
<td>rock climbing</td>
</tr>
<tr>
<td>citrus groves</td>
<td>san francisco</td>
</tr>
<tr>
<td>classical guitar</td>
<td>shakespeare</td>
</tr>
<tr>
<td>computer vision</td>
<td>stamp collecting</td>
</tr>
<tr>
<td>cruises</td>
<td>sushi</td>
</tr>
<tr>
<td>death valley</td>
<td>table tennis</td>
</tr>
<tr>
<td>field hockey</td>
<td>telecommuting</td>
</tr>
<tr>
<td>gardening</td>
<td>vintage cars</td>
</tr>
<tr>
<td>graphic design</td>
<td>volcano</td>
</tr>
<tr>
<td>gulf war</td>
<td>zen buddhism</td>
</tr>
<tr>
<td>hiv</td>
<td>zener</td>
</tr>
</tbody>
</table>

• Are these queries representative?
To compare our query-sensitive approach to ordinary PageRank, we conducted a user study. We randomly selected 10 queries from our test set for the study, and found 5 volunteers. For each query, the volunteer was shown 2 result rankings; one consisted of the top 10 results satisfying the query, when these results were ranked with the unbiased PageRank vector, and the other consisted of the top 10 results for the query when the results were ranked with the composite $s_{qd}$ score. The volunteer was asked to select all URLs which were “relevant” to the query, in their opinion. Furthermore, they were asked to say which of the two rankings was “better” overall, in their opinion. They were not told anything about how either of the rankings was generated. The rankings induced by the topic-sensitive PageRank score $s_{qd}$ were significantly preferred by our test group. Let a URL be considered relevant if at least 3 of the 5 volunteers selected it as relevant for the query. The precision then is the fraction of the top 10 URLs that are deemed relevant. The precision of the two ranking techniques for each test query is shown in Figure 1. The average precision for the rankings induced by the topic-sensitive PageRank scores is substantially higher than that of the unbiased PageRank scores. Furthermore, as shown in Table 7, for nearly all queries, a majority of the users preferred the rankings induced by the topic-sensitive PageRank scores. These results suggest that the effectiveness of a query-result scoring function can be improved by the use of a topic-sensitive PageRank scheme in place of a generic PageRank scheme.
Questions

[Haveliwala 2002]

- What happens for more obscure queries?
- What happens with ambiguous queries?
Questions
[Haveliwala 2002]

- Where do we get the topic-based vectors, $p$?
Questions
[Haveliwala 2002]

- Where do we get the topic-based vectors, $p$?
  - Open Directory Project
  - Wikipedia
  - automatic classification
  - browsing history
Questions
[Haveliwala 2002]

- Where do we get the topic-based vectors, $p$?
  - Open Directory Project
  - Wikipedia
  - automatic classification
  - browsing history

- Are there different personalization models?
News Personalization
News Personalization

- Assumption: People want to read items which are similar to what they have been historically interested in.
  - Is this accurate?
  - Is this desirable?
- Issues
  - scalability: must be able to deal with a large number of users.
  - dynamics: item set and relevance is constantly changing.
  - noise: click-based relevance signal is very noisy.
News Personalization

Related IR Work

- Information filtering [Belkin and Croft 1992]
  - user provide the system with a description of relevant information to track, the system filters relevant documents for presentation.
  - very important task in government information analysis.
- Topic detection and tracking [Allan 2002]
  - subtasks include detecting new topics (quickly), clustering, and filtering.
  - These methods are strictly content-based and do not assume the presence of a large number of users.
News Personalization

Problem Statement

\[ C_{ij} = \begin{cases} 
1 & \text{user } i \text{ clicked on document } j \\
0 & \text{otherwise} 
\end{cases} \]

- \( C^* \) \( N \times M \) matrix of clicks (‘ideal’)
- \( C \) \( N \times M \) matrix of observed clicks (‘training’)
- \( \tilde{C} \) \( N \times M \) matrix of estimated clicks (‘predicted’)
- \( N \) number of users
- \( M \) number of items
News Personalization

Problem Statement

\[ \text{precision} = \sum_{ij} \frac{(\tilde{C}_{ij} \land C^*_{ij})}{\sum_j \tilde{C}_{ij}} \]

\[ \text{recall} = \sum_{ij} \frac{(\tilde{C}_{ij} \land C^*_{ij})}{\sum_j C^*_{ij}} \]

(subject to efficiency constraints)
News Personalization

Problem Statement

precision

\[
\text{precision} = \frac{\sum_{ij} (\tilde{C}_{ij} \land C^*_{ij})}{\sum_{ij} \tilde{C}_{ij}}
\]

\[
\text{precision} = \frac{1}{N} \sum_i \frac{\sum_j (\tilde{C}_{ij} \land C^*_{ij})}{\sum_j \tilde{C}_{ij}}
\]

recall

\[
\text{recall} = \frac{\sum_{ij} (\tilde{C}_{ij} \land C^*_{ij})}{\sum_{ij} C^*_{ij}}
\]

\[
\text{recall} = \frac{1}{N} \sum_i \frac{\sum_j (\tilde{C}_{ij} \land C^*_{ij})}{\sum_j C^*_{ij}}
\]

what is the difference?
News Personalization

Approaches

\[ \tilde{C}_{ij} = \sum_{i \neq i'} w_{ii'} C_{i'j} \quad \text{memory-based} \]

\[ \tilde{C}_{ij} = \sum_{c \in \{ c | i \in c \}} w_{ic} \sum_{i' \in c} C_{i'j} \quad \text{model-based} \]

\( w_{ii'} \) affinity between users \( i \) and \( i' \)

\( w_{ic} \) affinity between users \( i \) and cluster \( c \)
News Personalization
Co-Visitation Approach
News Personalization

Co-Visitation Approach

\[ \tilde{C}_{ij} = \sum_{j'} \sum_{i'} C_{ij'} C_{i'j} \]
News Personalization
Co-Visitation Approach

\[ \tilde{C}_{ij} = \sum_{j'} \sum_{i'} C_{ij'} C_{i'j'} \]

\[ = \sum \sum \sum C_{ij'} C_{i'j'} C_{i'j} \]
News Personalization

Co-Visitation Approach

\[ \tilde{C}_{ij} = \sum_{j'} \sum_{i'} C_{ij'} C_{i'j} \]

\[ = \sum_{j'} \sum_{i'} C_{ij'} C_{i'j} C_{i'j} \]

\[ = \sum_{i'} C_{i'j} \sum_{j'} C_{i'j} C_{ij'} \]
News Personalization

Co-Visitation Approach

\[ \tilde{C}_{ij} = \sum_{j'} C_{ij'} \sum_{i'} C_{i'j'} C_{i'j} \]

\[ = \sum_{j'} \sum_{i'} C_{ij'} C_{i'j'} C_{i'j} \]

\[ = \sum_{i'} C_{i'j} \sum_{j'} C_{i'j'} C_{ij} \]

\[ = \sum_{i'} C_{i'j} \sum_{j'} C_{i'j'} C_{ij'} \]
News Personalization
Cluster-Based Approach

- **Locality-Sensitive Hashing**: hash the items viewed so that users in the same bucket tend to be similar.
  - very fast
  - can control precision
- **Probabilistic Latent Semantic Indexing**: cluster users according to their item visitation.
  - richer model with many extensions
  - more expensive
News Personalization

Experiments

- **Data**
  - MovieLens: public collaborative filtering dataset
  - NewsSmall: small sample of news records
  - NewsBig: larger sample of news records

- **Evaluation**
  - precision and recall curves
Experiments

Why is Correlation omitted from this plot?
Experiments

What should we infer from this plot?
Questions

- What other domains can we use this structure for?
Questions

- What other domains can we use this structure for?
- Do we need personalized news or should it be curated?
Questions

- What other domains can we use this structure for?
- Do we need personalized news or should it be curated?
- How could we combine this with the previous paper?
Deciding When To Personalize
### Should We Always Personalize?

<table>
<thead>
<tr>
<th>Query</th>
<th>Preferred by Majority</th>
</tr>
</thead>
<tbody>
<tr>
<td>alcoholism</td>
<td>TOPICSENSITIVE</td>
</tr>
<tr>
<td>bicycling</td>
<td>TOPICSENSITIVE</td>
</tr>
<tr>
<td>citrus groves</td>
<td>TOPICSENSITIVE</td>
</tr>
<tr>
<td>computer vision</td>
<td>TOPICSENSITIVE</td>
</tr>
<tr>
<td>death valley</td>
<td>TOPICSENSITIVE</td>
</tr>
<tr>
<td>graphic design</td>
<td>TOPICSENSITIVE</td>
</tr>
<tr>
<td>gulf war</td>
<td>TOPICSENSITIVE</td>
</tr>
<tr>
<td>hiv</td>
<td>NOBIAS</td>
</tr>
<tr>
<td>shakespeare</td>
<td>NEITHER</td>
</tr>
<tr>
<td>table tennis</td>
<td>TOPICSENSITIVE</td>
</tr>
</tbody>
</table>
Detecting When To Personalize

- Assumption: Disparate user intent suggests an opportunity to personalize.
  - Is this always true?
  - How would this affect system design?
- Approach: Study log data to find signals which correlate with multiple-intent queries (aka ‘potential for personalization’).
  - Gather ground truth data for queries which have multiple intents.
  - Model the relationship between observed data and the ground truth.
Detecting When To Personalize

Ground Truth

- Explicit
  - Relevance judgment for several individuals.
  - Potential for personalization: difference between upper bound on personalization performance and personalization performance with a user-agnostic ranker.

- Implicit
  - Record clicked URLs.
  - Measure the entropy of the clicked documents.
Relationship Between Explicit and Implicit Targets

Figure 3. Potential for personalization curve as a function of click entropy. For low click entropy, there is almost no potential for personalization, while for high click entropy there is a lot.
# Relationship Between Click Entropy and Number of Clicks

<table>
<thead>
<tr>
<th>Clicks/User</th>
<th>Low</th>
<th>Mid</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low</strong></td>
<td><a href="http://www.schoolloop.com">www.schoolloop.com</a></td>
<td>fox news network</td>
<td>ecw</td>
</tr>
<tr>
<td></td>
<td>usps.gov</td>
<td>ontario airport</td>
<td>fcc</td>
</tr>
<tr>
<td></td>
<td>men’s health magazine</td>
<td>wvu</td>
<td>arrow</td>
</tr>
<tr>
<td></td>
<td>espn2</td>
<td>larry king</td>
<td>internet explorer update</td>
</tr>
<tr>
<td><strong>Mid</strong></td>
<td>corvette america</td>
<td>michigan state football</td>
<td>toyota camry</td>
</tr>
<tr>
<td></td>
<td>cleartype</td>
<td>alaska cruise</td>
<td>rachel ray recipes</td>
</tr>
<tr>
<td></td>
<td>petfinder.org</td>
<td>trivia quiz</td>
<td>bruce springsteen lyrics</td>
</tr>
<tr>
<td></td>
<td>pfchang</td>
<td>knee injury</td>
<td>stress hormones</td>
</tr>
<tr>
<td><strong>High</strong></td>
<td>(no queries)</td>
<td>restaurant guide</td>
<td>first aid</td>
</tr>
<tr>
<td></td>
<td></td>
<td>famous poems</td>
<td>hand foot mouth disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td>calculate bmi</td>
<td>cupcake recipes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>woodrow wilson</td>
<td>house spiders</td>
</tr>
</tbody>
</table>
## Signals Believed to be Correlated with Multi-Intent Queries

<table>
<thead>
<tr>
<th>Information</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Query</td>
<td>Query length (chars, words)</td>
</tr>
<tr>
<td></td>
<td>Contains location</td>
</tr>
<tr>
<td></td>
<td>Contains URL fragment</td>
</tr>
<tr>
<td></td>
<td>Contains advanced operator(s)</td>
</tr>
<tr>
<td></td>
<td>Time of day issued</td>
</tr>
<tr>
<td></td>
<td>Issued during work hours</td>
</tr>
<tr>
<td></td>
<td>Number of results</td>
</tr>
<tr>
<td></td>
<td># of query suggestions offered</td>
</tr>
<tr>
<td></td>
<td># of ads (mainline and sidebar)</td>
</tr>
<tr>
<td></td>
<td>Has a definitive result</td>
</tr>
<tr>
<td>Results</td>
<td>Query clarity</td>
</tr>
<tr>
<td></td>
<td>ODP category entropy</td>
</tr>
<tr>
<td></td>
<td># of ODP categories</td>
</tr>
<tr>
<td></td>
<td># of distinct ODP categories</td>
</tr>
<tr>
<td></td>
<td># of URLs matching ODP</td>
</tr>
<tr>
<td></td>
<td>Portion of results non-html</td>
</tr>
<tr>
<td></td>
<td>Portion that are “.com”/”.edu”</td>
</tr>
<tr>
<td></td>
<td># of distinct domains</td>
</tr>
</tbody>
</table>
Results

- Relationship between individual signals and ambiguous intent.
  - signals based on history have the highest correlations with targets.
  - signals not based on history have weaker correlations with targets.
- Accuracy of ensembles of signals and ambiguity.
  - query features alone provide strong baseline.
  - result features provide marginal improvement.
  - combinations of query and result history provide the strongest performance.
Questions

- How should the system respond to the prediction of an ambiguous intent?
- Why not always personalize?